

1 **DISPLAY ANNOTATION AND LAYOUT PROCESSING**

2 FIELD OF THE INVENTION

3 The present invention relates to an information processing
4 method and an information processing system. More
5 particularly, the present invention relates to improved
6 provision of annotation and/or layout for display.

7 BACKGROUND

8 The use of the Internet became popular. As the role of the
9 Internet has been varied, variety of apparatuses for access
10 to the Internet becomes more diverse. Conventionally, a
11 computer system having a CRT (Cathode Ray Tube) with a
12 display area of about 12 to 20 inches, a liquid crystal
13 display or a plasma display device has been used as an
14 apparatus for connection to the Internet.

15 However, while taking into account cases wherein portability
16 is important, there has been a dramatic spread in the
17 integration of handy telephones, PDAs (Personal Digital
18 Assistants) and i-mode handy phones. These apparatuses are
19 generally having small display area. Further, since visually
20 impaired persons cannot confirm the output of computers by
21 observing display devices, the reading software, such as
22 speech browsers, has been developed. It is anticipated that
23 such reading software will eventually constitute a human

1 interface improvement, not only for visually impaired persons
2 but also for users who are unfamiliar with computers. Then,
3 this kind of software technique can contribute to and promote
4 the wider use of computer systems. In addition, for wearable
5 computers, since the areas of their display devices should
6 perforce be small, it is predicted that speech output will be
7 a primary or, at the least, an auxiliary output means.

8 In general, the designs of page layouts for web sites are
9 based on the assumption that the display devices of computer
10 systems will have 12 to 20 inch display areas. Furthermore,
11 for the output to display devices, it is premised that
12 displays will be used by persons with normal sight.
13 Specifically, the menu area (link information is embedded
14 there) of a site and an advertisement banner are ordinarily
15 arranged at the upper or left portion of a display area, and
16 the two-dimensional layout of the data is presented, so that
17 it can be easily viewed by users with normal sight. The
18 information inherent to a page commonly tends to be arranged
19 in the center or in the latter half of a page layout.

20 When a web page, designed for users with normal sight, or a
21 large screen device, is to be displayed on a PDA or a
22 portable telephone, or is to be output by a speech browser,
23 usually the information (frame information, an advertisement,
24 etc.) at the first of a page tends to be an obstacle. The
25 two-dimensional information, such as frame information or
26 advertisements, is effective and improves the usability for
27 users with normal sight and a large screen. However, for
28 users who operate small screen devices or employ speech

1 browsers, these secondary information becomes an obstacle to
2 find the most important information, such as the inherent
3 information of the page. Therefore, when a device having a
4 small screen or a speech browser is employed to output a page
5 file designed for a large screen, we have to provide some
6 method for accessing to important information easily.

7 Therefore, when a device having a small screen or a speech
8 browser is employed to output a page file designed for a
9 large screen, some means is required for rapidly accessing
10 the initially sought information. One well-known means is a
11 method that provides annotations for a page file. Annotation
12 is an additional data, such as the structure of a page file
13 and the importance level of each portion. Usually, the
14 annotation is written to an external file, and is used to
15 simplify page file accurately.

16 However, it is not easy to provide annotations for each page
17 file. Generally, while each page file is browsed and the
18 display is conformed, the importance level of the page file
19 should be determined and annotations should be provided.
20 These operations need be performed manually. Especially at a
21 news site or a database site, the annotator's workload to
22 prepare annotations is significantly increased because the
23 volume of available page files is large. In addition, when a
24 new file is to be generated by including date data in the URL
25 (Uniform Resource Locator), even if annotations have already
26 been provided to the site, additional annotations should be
27 prepared.

1 SUMMARY OF THE INVENTION

2 It is, therefore an aspect of the present invention to
3 provide methods, apparatus and systems for preparing
4 annotations for a page file. Thus, according to the present
5 invention, a example method is provided whereby a page group
6 employing the same layout is detected in accordance with the
7 tag structure of a document, such as an HTML (HyperText
8 Markup Language) document. Then, annotations are shared
9 among these pages. At a site designated by a user, the layout
10 structure of the contents is analyzed, and tags (hereinafter
11 referred to as layout tags) are enumerated that are factors
12 referred to when determining a layout. At the same time, in
13 order to clearly identify the structure of the layout tags in
14 a document, such as an HTML document, the layout tags are
15 written in a structural descriptive form that employs a style
16 for the designation of positions on the page, i.e., an XPath,
17 an XPointer or a tree format. Further, the characteristic
18 values of the layout tags (structural descriptive forms) are
19 acquired. Then, based on the obtained data, the distance
20 between the pages is calculated. Based on the calculated
21 distance, a group of pages using the same layout and a group
22 of pages sharing part of the layout are automatically
23 detected and presented to a user. When the user adds an
24 annotation for one representative page of a page group, a
25 corresponding annotation is added to [generally to all] pages
26 in the group that employ the same layout. When there are
27 pages that share the layout, first, an annotation is added to

1 the portion used in common, and then, annotations are added
2 to the portions that are individually held by individual page
3 groups. In this fashion, an efficient annotation provision
4 can be provided.

5 Further, in this invention, when a user additionally performs
6 a correction to divide or unify the presented page groups,
7 the results can be employed to correct the distance
8 calculation expression. As a result, the accuracy in the
9 following page group division can be improved.

10 BRIEF DESCRIPTION OF THE DRAWINGS

11 These and other aspects, features, and advantages of the
12 present invention will become apparent upon further
13 consideration of the following detailed description of the
14 invention when read in conjunction with the drawing figures,
15 in which:

16 Fig. 1 is a block diagram showing an example information
17 processing system according to one embodiment of the present
18 invention;

19 Fig. 2 is a block diagram showing an example structure of an
20 HTML file analysis module;

21 Fig. 3 is a diagram showing a URL and layout tags, and
22 characteristic values that are related to the URL.

1 Fig. 4 is a block diagram showing an example structure for a
2 page group detection module;

3 Fig. 5 is a diagram showing a screen obtained by browsing
4 example page files that fall into the same layout group;

5 Fig. 6 is a diagram showing a screen obtained by browsing
6 another example page file that falls in the same layout
7 group;

8 Fig. 7 is a diagram showing a screen obtained by browsing
9 example page files that do not fall into the same layout
10 group;

11 Fig. 8 is a flowchart showing the annotation addition
12 processing;

13 Fig. 9 is a flowchart showing the processing for adding an
14 annotation to a page group for which a temporary layout ID
15 was provided; and

16 Fig. 10 is a flowchart showing the processing for adding an
17 annotation to a layout sharing group.

18 DESCRIPTION OF THE SYMBOLS

19 1: Information processing system
20 2: Database

1 3: Page acquisition module
2 4: HTML file analysis module
3 5: Page group detection module
4 6: Annotation addition module
5 7: Correction module for the function of distance
6 calculation
7 8: Web server
8 9: Objective URL list
9 10: Annotation addition
10 20: HTML parser
11 21: Layout tag listing module
12 22: Characteristic value acquisition module
13 41: Inter-page distance calculation module
14 42: Layout group determination module
15 43: Representative value calculation module
16 44: Inter-layout distance calculation module
17 45: Layout sharing group determination module

18 DESCRIPTION OF THE INVENTION

19 The present invention provides methods, apparatus and systems
20 whereby a page group employing the same layout is detected in
21 accordance with the tag structure of a document, such as an
22 HTML document. Then, annotations are shared among these
23 pages. At a site designated by a user, the layout structure
24 of the contents is analyzed, and tags are enumerated that are
25 factors referred to when determining a layout. At the same
26 time, in order to clearly identify the structure of the

1 layout tags in a document, such as an HTML document, the
2 layout tags are written in a structural descriptive form that
3 employs a style for the designation of positions on the page,
4 i.e., an XPath, an XPointer or a tree format. Further, the
5 characteristic values of the layout tags (structural
6 descriptive forms) are acquired. Then, based on the obtained
7 data, the distance between the pages is calculated. Based on
8 the calculated distance, a group of pages using the same
9 layout and a group of pages sharing part of the layout are
10 automatically detected and presented to a user. When the
11 user adds an annotation for one representative page of a page
12 group, a corresponding annotation is added to [generally to
13 all] pages in the group that employ the same layout. When
14 there are pages that share the layout, first, an annotation
15 is added to the portion used in common, and then, annotations
16 are added to the portions that are individually held by
17 individual page groups. In this fashion, an efficient
18 annotation provision can be provided.

19 Further, in this invention, when a user additionally performs
20 a correction to divide or unify the presented page groups,
21 the results can be employed to correct the distance
22 calculation expression. As a result, the accuracy in the
23 following page group division can be improved.

24 An example embodiment of the present invention will now be
25 described in detail while referring to the accompanying
26 drawings. It should be noted, however, that the present
27 invention can be implemented by various other embodiments,
28 and is not limited to this embodiment. Further, throughout

1 this embodiment, the same reference numerals are used to
2 denote corresponding or identical components.

3 In the embodiment, mainly, a method or a system will be
4 explained. However, as will be apparent to one having
5 ordinary skill in the art, the present invention can be
6 implemented not only as a method and a system, but also as a
7 computer-readable program, or as a storage medium on which
8 such a program is stored. Therefore, the present invention
9 can be provided as hardware, software or a combination of
10 hardware and software. An example storage medium on which
11 the program can be recorded is an arbitrary computer-readable
12 storage medium, such as a hard disk, a CD-ROM, an optical
13 storage device or a magnetic storage device.

14 In the following embodiment, a common computer system can be
15 employed. The computer system used for this embodiment
16 comprises a central processing unit (CPU), a main memory
17 (RAM) and a nonvolatile memory (ROM), [generally to all] of
18 which are interconnected by a bus. In addition, a
19 co-processor, an image accelerator, a cache memory and an
20 input/output controller (I/O) may be connected to the bus.
21 Further, an external storage device, a data input device, a
22 display device and a communication controller are also
23 connected to the bus via an appropriate interface, as are the
24 hardware resources generally provided for a computer system.
25 An example external storage device is a hard disk drive;
26 however, a device such as a magneto-optical storage device,
27 an optical storage device or a semiconductor storage device,
28 such as a flash memory, can also be employed as an external

1 storage device. As the data input device, a device such as a
2 keyboard, a pointing device, such as a mouse, a pen input
3 device or a tablet can be employed. The data input device
4 also includes an image reader, such as a scanner, or a speech
5 input device. An example display device can be a CRT, a
6 liquid crystal display device or a plasma display device.
7 Furthermore, the computer system includes an arbitrary
8 computer, such as a personal computer, a workstation or main
9 frame computer.

10 Fig. 1 is a block diagram showing an example information
11 processing system according to one embodiment of the present
12 invention. An information processing system 1 of this
13 embodiment comprises a database 2, a page acquisition module
14 3, an HTML file analysis module 4, a page group detection
15 module 5, an annotation addition module 6 and a correction
16 module 7 for the function of distance calculation.

17 The database 2 is used to record data generated by modules
18 that will be described later and a page file (also called an
19 HTML file) obtained from a web server 8. The database 2 is
20 constituted by a storage device, such as a hard disk drive,
21 that is internally provided for the information processing
22 system 1 of this embodiment and software for controlling the
23 input/output of data. However, the database 2 is not
24 necessarily provided inside the information processing system
25 1, and may be an external file as designated by a URL.
26 Further, the database 2 need not be intensively managed, and
27 may be recorded and managed in a distributed manner. That
28 is, so long as the input/output of necessary data can be

1 carried out by appropriate address designation means, the
2 database 2 of this embodiment can be constituted, regardless
3 of the type of physical storage device or its location.

4

5 The page acquisition module 3 receives an objective URL list
6 9 from a user, and obtains the contents of the associated URL
7 from the web server 8. For example, HTTP (HyperText Transfer
8 Protocol) is used for an acquisition request, and the
9 obtained HTML file (page file) will be recorded in the
10 database 2.

11 First, the page acquisition module 3 obtains the page file of
12 the objective URL list 9. Then, the URLs (e.g., obtained
13 from the href attribute of <a> tag) included in a page at the
14 objective URL are enumerated, and from among these URLs, only
15 a URL included in a range designated by a user is selected
16 and is added to the URL list 9. Following this, the pages on
17 the URL list 9 are sequentially obtained, and as the page
18 files are obtained, they are recorded in the database 2.

19 When URLs that are related to the associated URL that is
20 obtained are included, the same process is recurrently
21 performed for these associated URLs. In this manner, pages
22 linked in the site can be obtained. Meanwhile, a double
23 registration should not be performed for a URL that has
24 already appeared on the URL list 9. The URL list 9 is also
25 recorded in the database 2.

26 The HTML file analysis module 4 analyzes the page files
27 obtained by the page acquisition module 3 in order to list
28 the layout tags that affect the page layout and to obtain the

1 characteristic values of the layout tags.

2 Fig. 2 is a block diagram showing an example configuration
3 for the HTML file analysis module 4. The HTML file analysis
4 module 4 includes an HTML parser 20, a layout tag listing
5 module 21 and a characteristic value acquisition module 22.

6 The HTML parser 20 analyzes the HTML file obtained by the
7 page acquisition module 3, and converts the HTML file into a
8 tag structure description form, such as a DOM tree.

9 The layout tag listing module 21 employs the structural
10 descriptive form to list, from the obtained tag structure,
11 the tags (layout tags) that affect the layout structure.
12 Example layout tags can be "table", "tbody", "tr", "td", "th"
13 and "hr". The style, such as the XPath or XPointer, for
14 designating the position on a page, or the tree format can be
15 employed as the structural descriptive form.

16 The characteristic value acquisition module 22 correlates,
17 with the structure description form, the characteristic
18 values of the attributes of the listed layout tags and
19 elements that are included in the sub-trees of the layout
20 tags. The following attributes and elements can be employed
21 as the characteristic values. For layout tag "table", there
22 are the attributes "align", "bgcolor", "border",
23 "cellpadding", "cellspacing" and "width". For layout tag
24 "tbody", there are the attributes "align" and "valign". For
25 layout tag "tr", there are the attributes "align", "bgcolor"
26 and "valign". For layout tag "td" or "th", there are the

1 attributes "align", "bgcolor", "colspan", "height",
2 "rowspan", "valign" and "width" and the presence/absence of
3 the element, such as text or an image, and the size of the
4 element. And for layout tag "hr", there are the attributes
5 "align", "width", "size" and "noshade".

6 The HTML file analysis module 4 correlates the layout tags
7 having the structural descriptive form and the correlated
8 characteristic values with the URLs of the URL list. The
9 HTML file analysis module 4 then records the layout tags and
10 the characteristic values in the database 2.

11 Fig. 3 is a diagram showing a URL on the URL list and the
12 layout tags and the characteristic values that are correlated
13 with the URL. For example, URL

14 "http://www.ibm.com/index.html"

15 includes layout tags

16 "/html[1]/body[1]/table[1]", and

17 "/html[1]/body[1]/table[1]/tr[1]/td[1]",

18 which are written in the structure description form (XPath in
19 this case). Characteristic values "width=200, bgcolor=blue,
20 . . ." are correlated with "/html[1]/body[1]/table[1]", while
21 characteristic value "bgcolor=red, . . ." is correlated with
22 "/html[1]/body[1]/table[1]/tr[1]/td[1]".

23 The page group detection module 5 calculates an inter-page

1 distance by using the layout tags and the characteristic
2 values that are obtained by the HTML file analysis module 4.
3 With this function, the page group detection module 5
4 extracts, as a layout group, a group of pages having the same
5 or similar layout structure. In addition, the page group
6 detection module 5 calculates, for one part of the area of
7 the page file, a layout having a layout structure used in
8 common by another page file, and extracts these page files as
9 a layout sharing group.

10 Fig. 4 is a block diagram showing an example structure for
11 the page group detection module 5. The page group detection
12 module 5 includes an inter-page distance calculation module
13 41, a layout group determination module 42, a representative
14 value of layout group calculation module 43, an inter-layout
15 distance calculation module 44, and a layout sharing group
16 determination module 45.

17 The inter-page distance calculation module 41 employs a
18 characteristic value correlated with the layout tag to
19 calculate a distance between a page file including the layout
20 tag and another page file. The layout group determination
21 module 42 extracts, as a layout group, page files for which
22 the inter-page distance calculated by the inter-page distance
23 calculation module 41 falls within a predetermined range.
24 The representative value calculation module 43 calculates a
25 representative value for page file groups that are layout
26 groups and have the same or similar layout structure. The
27 inter-layout distance calculation module 44 calculates the
28 distance between layout groups. The layout sharing group

1 determination module 45 determines whether part of page files
2 in a layout group includes the same or similar layout
3 structure used in common by page files in other layout
4 groups. When there is a layout used in common, the page
5 files in the layout groups are extracted as layout sharing
6 groups.

7 There are several methods that can be used for calculating
8 the distance between pages. For this embodiment, an
9 explanation will now be given for a method whereby the layout
10 tags and their characteristic values are weighted, and the
11 total of the distances between these tags is defined as an
12 inter-page distance. Assuming that A and B denote sets of
13 structural descriptive forms for layout tags included on two
14 target pages for distance calculation, the inter-page
15 distance D is represented by the following equation.

$$16 \quad D = \sum d_i(T_i)$$

17 where T_i denotes the i -th element of the layout tag that
18 satisfies $A \cup B$, and d_i denotes the distance function of
19 layout tag T_i . It should be noted that i satisfies $1 \leq i \leq$
20 (the total number of layout tags that satisfy $A \cup B$).

21 The distance function d_i is a function of the layout tag T_i ,
22 and when $T_i \in (A \cap B)$,

$$23 \quad d_i(T_i) = W_i * \sum W_{cij} * (f_i(C_{Aij}, C_{Bij})),$$

24 while in other cases,

$$25 \quad d_i(T_i) = W_i * L_i,$$

1 where W_i denotes a weighting coefficient for the layout tag
2 T_i , and "1", for example, can be employed. C_{ij} denotes the
3 value of a characteristic value j for the layout tag T_i . W_{cij}
4 denotes the weighting coefficient for the characteristic
5 value C_{ij} of the layout tag T_i , and "1", for example. f_i
6 denotes a function that represents the distance between the
7 characteristic values, while a function for returning a "0"
8 when the characteristic values are the same and for returning
9 a "1" when they differ can be employed. L_i denotes a
10 distance constant when the layout tag T_i is present only on
11 one page, and, for example, $L_i = 5$ can be employed.

12 The inter-page distance calculation module 41 calculates the
13 inter-page distance D using the above method, and the layout
14 group determination module 42 employs the inter-page distance
15 D to group the same or similar layouts. A method, such as
16 clustering, can be employed for this determination means, and
17 the inter-layout distance D of equal to or smaller than
18 threshold value e.g. 10 can be employed as the reference for
19 determination of the similarity range.

20 An example page file that constitutes the thus generated
21 layout group is shown in Figs. 5A and 5B. Fig. 5A is a
22 diagram showing a screen presented by browsing a specific
23 page file, and Fig. 5B is a diagram showing a screen
24 presented by browsing a second page file. The distance
25 between these pages obtained by the above method is "0" in
26 this case. That is, in the structure of the page layout, the
27 layout tags and characteristic values are the same for the
28 file in Fig. 5A and the file in Fig. 5B. Thus, these two

1 page files fall into the same layout group. Naturally,
2 however, contents irrelevant to the layout structure (the
3 contents of individual table elements) differ.

4 Another example of page files in the same layout group is
5 shown in Figs. 6A and 6B. Fig. 6A is a diagram showing the
6 screen obtained by browsing a specific page file, and Fig. 6B
7 is a diagram showing the screen obtained by browsing a second
8 page file. The inter-page distance obtained by the above
9 method is "3" in this case, and both of the page files have
10 the same layout tag structure. However, the layout tags
11 related to the layouts for portions indicated by arrows have
12 different characteristic values (display colors in this
13 example). In this example, an inter-page distance of "3" is
14 obtained because of this difference. However, since the
15 inter-page distance does not exceed "10", it is ascertained
16 that the page files are similar and fall into the same layout
17 group.

18
19 Figs. 7A and 7B are diagrams showing examples of screens of
20 page files that do not fall into the same layout group. The
21 page files in Figs. 7A and 7B are displayed by browsing, as
22 are those in Figs. 5 and 6. In this case, the layout tag
23 structures are the same. However, the characteristic values
24 of the layout tags of the two page files differ greatly, and
25 it is ascertained that the page files have different layouts.
26 For example, for layout tag "td" at the portions indicated by
27 arrows, in Fig. 7A characters are arranged by setting
28 "width", while in Fig. 7B an image is simply located.
29 Further, in Fig. 7A "bgcolor" is set for the layout tag "tr",

1 while in Fig. 7B "bgcolor" is not set. Because of these
2 differences, an inter-page distance D of "14" is obtained,
3 and the page files fall into different layout groups.

4 Through this processing, the grouping of same or similar page
5 files is accomplished and the obtained layout groups are
6 recorded in the database 2.

7 An explanation will now be given for the processing for
8 extracting a layout sharing group having the same layout of
9 one part of a page file. For each layout group obtained by
10 the above method, the representative value calculation module
11 43 calculates the representative values of the layout group
12 based on the layout tags and the characteristic values.
13 First, the representative value calculation module 43 obtains
14 a layout tag that is representative of the layout group. The
15 method for obtaining a representative tag can be a method for
16 calculating a set of sums or a set of products of the layout
17 tags included in the page files of the layout group. As
18 other methods, there are a method can be a method for
19 obtaining a set of layout tags such that the number of page
20 files having a specific layout tag exceeds a threshold value,
21 and an arbitrary method for determining a tag representative
22 of the layout tags for the layout group. Subsequently, the
23 representative value calculation module 43 determines the
24 characteristic values of the selected layout tags. A method
25 for determining the characteristic values can be one whereby
26 a decision is obtained based on a majority or an average of
27 the characteristic values of the page files in the layout
28 group.

1 The inter-layout distance calculation module 44 calculates
2 the distance between the layout groups by using the
3 representative values for the individual layout groups
4 obtained by the representative value calculation module 43.

5 Several methods are available for calculating the distance
6 between layout groups. In this embodiment, an explanation
7 will be given for a method for performing weighting for a
8 layout tag and its characteristic value, and for obtaining,
9 as an inter-layout distance, the sum of distances between
10 tags. When A' and B' denote sets of representative tags
11 belonging to two layout groups between which the distance is
12 to be calculated, the inter-layout distance D' is represented
13 by the following equation,

$$14 \quad D' = \sum d_i'(T_i)$$

15 where T_i denotes the i -th element of layout tags that satisfy
16 $A' \cup B'$, and d_i' denotes the distance function for the layout
17 tag T_i . It should be noted that i is $1 \leq i \leq$ (the total of
18 the layout tags that satisfy $A' \cup B'$).

19 The distance function d_i' is the function of the layout tag
20 T_i , and when $T_i \in (A' \cap B')$,

$$21 \quad d_i'(T_i) = W_i' * (M_i + \sum W_{Cij}' * (f_i'(C_{Aij}, C_{Bij})),$$

22 is established, whereas in another case,

1 $d_i'(T_i) = W_i' * L_i'$.
2 W_i' denotes the weighting coefficient of the layout tag T_i ,
3 and is, for example, "1". C_{ij}' denotes the characteristic
4 value j of the layout tag T_i . $W_{C_{ij}}'$ denotes the weighting
5 coefficient of the characteristic value C_{ij} of the layout tag
6 T_i , and is, for example, "1". f_i' denotes a function that
7 represents the distance between characteristic values. For
8 f_i' , a function can be employed that returns a "0" when the
9 characteristic values are the same or that returns a "1" when
10 the characteristic values differ. M_i denotes the distance
11 constant when the layout tag T_i is present in both of the
12 layout groups. L_i' denotes the distance constant when the
13 layout tag T_i is present in only one. In this manner, the
14 distance D' , which separates the layout groups, can be
15 obtained.

16 The layout sharing group determination module 45 employs the
17 inter-layout distance D' , which is supplied by the
18 inter-layout distance calculation module 44, to group page
19 files using a method such as clustering. Then, those page
20 groups (layout sharing groups) that are assumed to share a
21 part of the layout are enumerated. It should be noted that
22 inherent layout IDs are allocated for the layout groups or
23 the layout sharing groups.

24 In response to an annotation addition request 10 issued by a
25 user, the annotation addition module 6 adds an annotation to
26 each group. To add an annotation to an entire layout group,
27 the annotation addition module 6 correlates the annotation

1 with an inherent layout ID allocated for the layout group.

2 For the addition of the annotation, a page group (a layout
3 group or a layout sharing group) detected by the page group
4 detection module 5 is presented to the user. At this time,
5 the relationship of the sharing of the layout is depicted
6 using a graphical method, e.g. tree graph, it can be easily
7 understood by the user.

8 Sequentially, the user selects a page from the presented page
9 group, and adds the annotation to the selected page. Then,
10 the annotation is stored in the database 2, correlated with
11 the layout ID of the pertinent page. When a layout sharing
12 group is present, the annotation added to the tag structure
13 that is stored in common (hereinafter referred to as a
14 sharing layout) is copied to and stored in correlation with
15 the layout ID of each element of the layout sharing group.

16 When the user selects a page for which the annotation has
17 already been added to the sharing layout portion, the sharing
18 layout portion is highlighted and presented to the user, so
19 that the annotation information can be referred to.
20 Therefore, the user need only add the annotation to the
21 portion that the layout group independently stores, and can
22 add the annotation for the entire page.

23 When the user divides or unifies layout groups or separates
24 members of a sharing relationship, the correction module 7
25 for the function of distance calculation corrects the
26 parameters used for distance calculation, so that they

1 reflect the division or unification or the separation.

2 When the user corrects the presented page group, for example,
3 by dividing or unifying it, the inter-page distance
4 calculation expression is corrected using the correction
5 results, and the accuracy of the division of a page group can
6 thereafter be increased. To make the correction, various
7 methods can be employed. For this embodiment, an explanation
8 that will now be given describes a method used to change the
9 inter-page distance calculation expression by changing the
10 weighting provided for the layout tag and the characteristic
11 value.

12
13 When the division of a layout group is instructed, in the
14 groups obtained by the division, different layout tags and
15 characteristic values are employed. The inter-page distance
16 calculation expression is changed by increasing the weighting
17 for layout tags and for characteristic values, and during the
18 following page group detection process, these layout groups
19 are detected as different groups. It should be noted that
20 the weighting may be reduced for layout tags, which are
21 matched for the groups obtained by the division, and for
22 characteristic values.

23 When the merging (unification) of layout groups is
24 instructed, contrary to what is described above, the
25 weighting for the layout tags and the characteristic values
26 is reduced. And the calculation expression is changed, so
27 that during the following page group detection process and
28 the layout sharing determination process, these layout groups

1 are determined to be members of the same page group or layout
2 sharing group. It should be noted that in a merged group the
3 weighting of layout tags and characteristic values that match
4 may be increased.

5 When the user adds a correction, such as the cancellation
6 (separation) of a layout sharing relationship, similarly, the
7 layout tags and characteristic values that differ between the
8 representative values for the layout groups are employed.

9 The inter-layout distance calculation expression is corrected
10 by changing the weighting provided for these layout tags and
11 characteristic values. As a result, the accuracy attained in
12 the determination of the layout sharing can thereafter be
13 increased.

14 An overview of the information processing system of this
15 embodiment has been given. Now, an explanation will be
16 presented for an annotation addition method that uses this
17 system. First, a user designates the URL of an object site
18 and the condition (the directory or the updating date) of an
19 object to which an annotation is to be added. Then, during
20 the processing performed by the information processing
21 system, the page acquisition module 3 obtains an object HTML
22 file, the HTML file analysis module 4 analyzes the page file,
23 and the page group detection module 5 detects a layout group
24 and a layout sharing group.

25 Following this, the page groups (layout groups) that are
26 assumed to have the same layout are presented to the user in
27 an arbitrary order, such as the descending order of the

1 number of page files in the page group. Then, a request is
2 issued for the addition of an annotation to an arbitrary page
3 (page file) in the page group.

4 Fig. 8 is a flowchart showing the annotation addition
5 processing. First, as is described above, layout groups
6 (page groups) are obtained from the database 2 and are
7 presented to the user (step 50). Then, a check is performed
8 to determine whether an annotation has been added to
9 [generally to all] the layout groups (step 51). When an
10 annotation has been added to the layout groups, the
11 processing is terminated (step 52). But when an annotation
12 has not yet been added to one or more layout groups, program
13 control shifted to step 53. At step 53, an arbitrary layout
14 group (page group) is selected, and a layout ID(1) is
15 selected for correlation with the page group.

16 Then, an arbitrary page (page file) in the page group (layout
17 group) is selected by the user (step 54). Thereafter, at
18 step 55, the selected page file is presented to the user by
19 an appropriate browser, and the user, while watching the
20 display screen, adds an annotation. Specifically, the user
21 adds, for example, a link for jumping to a screen division
22 for a PDA or a small screen device, or to the content of a
23 speech browser. The layout ID(1) is then correlated with the
24 added annotation.

25 After the annotation has been provided, the number of
26 applicable pages in the page group is presented to permit the
27 user to select either to present the annotation provided for

1 the entire page group, or to apply the annotation for the
2 individual pages. That is, a check is performed to decide
3 whether it is possible to use the annotation for the entire
4 page group (layout group) (step 56). When the decision at
5 step 56 is 'Yes', the layout ID(1) is provided for [generally
6 to all] the page files in the page group (step 57), and
7 program control advances to step 58 for the provision of an
8 annotation for the layout sharing group.

9 When the decision at step 56 is 'No', a check is performed to
10 determine whether it is possible to add the annotation to
11 selected pages of the page group. At step 59 a check is
12 performed to confirm that [generally all] pages in the page
13 group have been processed. When the decision is 'No', one of
14 the remaining pages is selected (step 60).

15 A check is then performed to determine whether it is possible
16 to use the annotation for the selected page (step 61). When
17 it is determined the use of the annotation is possible (the
18 decision at step 61 is 'Yes'), the layout ID(1) is provided
19 for the selected page (step 62). When it is determined use
20 of the annotation is not possible (the decision at step 61 is
21 'No'), a temporary layout ID is provided for the selected
22 page (step 63). This temporary layout ID is a common ID
23 provided for pages for which the layout ID(1) can not be
24 used, and an identification ID for the performance of the
25 individual processes, as will be described later.

26 After the layout ID(1) or the temporary layout ID has been
27 provided, program control returns to step 59, and the

1 processing at step 59 and the following steps is repeated.
2 When it is ascertained at step 59 that [generally to all] the
3 pages in the page group have been processed, a check is
4 performed to determine whether a page is present for which
5 the temporary layout ID was provided (step 64). When the
6 decision is 'Yes', program control advances to a process
7 (step 65) for adding an annotation to a page group for which
8 a temporary layout ID was provided. When no pages remain for
9 which the temporary layout ID was provided, program control
10 advances to step 58.

11 Fig. 9 is a flowchart showing the processing for adding an
12 annotation to a page group for which the temporary layout ID
13 has been provided. When program control advances to step 65
14 in the flowchart in Fig. 8, the processing in Fig. 9 is
15 performed. First, an arbitrary page is selected from the
16 page group including pages for which the temporary layout ID
17 was provided (step 70), and a layout ID(2) is provided for
18 the selected page. Then, an annotation is added to the
19 selected page (step 71). The layout ID(2) is provided for
20 the annotation. A check is then performed to determine
21 whether the annotation can be added to [generally to all] the
22 pages in the page group that were provided the temporary
23 layout ID (step 72). When the decision is 'Yes', the layout
24 ID(2) is added to [generally to all] the pages of the page
25 group that were originally provided the temporary layout ID
26 (step 73). The inter-page distance calculation expression is
27 then corrected (step 74), and thereafter the processing is
28 terminated (step 75).

1 When the decision at step 72 is 'No' (when the annotation can
2 not be used for all the pages in the page group that were
3 provided the temporary layout ID), a check should be
4 performed to determine whether the annotation can be applied
5 for individual pages. At step 76, a check is performed to
6 determine whether it is confirmed that the annotation can be
7 added to [generally to all] the pages in the page group for
8 which the temporary layout ID was provided. When the
9 confirmation is not yet completed (the decision is 'No'), an
10 arbitrary page is selected from the page group (step 77), and
11 a check is performed to determine whether the application of
12 the annotation for the selected page is possible (step 78).
13 When the application is possible, the layout ID(2) is
14 provided for the selected page (step 79) and program control
15 returns to step 76. When, at step 78, the annotation can not
16 be applied, program control returns to step 76 without
17 performing any further processes (maintains the temporary
18 layout ID).

19 When the decision at step 76 is 'Yes' (the confirmation for
20 the pages has been completed), a check is performed to
21 determine whether there is a page for which the temporary
22 layout ID was provided (step 80). When there is no page for
23 which the temporary layout ID was provided (the decision is
24 'No'), program control is shifted to step 74, and the
25 inter-page distance calculation expression is corrected. The
26 processing is thereafter terminated (step 75). But when
27 there is a page for which the temporary layout ID is provided
28 (the decision at step 80 is 'Yes'), program control returns
29 to step 70 and the above processing is repeated.

1 Through this processing, [generally all] pages having the
2 temporary layout ID are processed and an appropriate
3 annotation is assigned to each of the pages of the target
4 page group (layout group). When different annotations are
5 provided for pages in the same layout group, at step 74 the
6 inter-page distance calculation expression is corrected.
7 Thus, through the calculation of the next inter-page
8 distance, the correction is reflected and the pertinent pages
9 are sorted into different layout groups.

10 The processing for adding an annotation to the layout sharing
11 group (step 58) will now be described. Fig. 10 is a
12 flowchart showing the processing for adding an annotation to
13 the layout sharing group. First, an arbitrary page group
14 (layout group) is selected from among the layout sharing
15 groups (step 81). Then, a check is performed to determine
16 whether there are multiple annotation choices to be added to
17 the sharing layout (step 82). Since a page group is divided
18 or different annotations are provided in the layout sharing
19 group, it is highly probable that multiple annotation choices
20 will be available for the layout sharing portion. In this
21 case, in the following process for adding an annotation to a
22 layout sharing group, annotation choices are presented in
23 order to permit a user to select one of them (step 83).
24 Then, a check is performed to determine whether the selected
25 annotation can be applied for the layout sharing portion
26 (step 84). When the application is possible, the annotation
27 to be added to the sharing portion is copied, and provision
28 of the annotation for portions other than the sharing portion

1 is requested (step 86). The above described method is used
2 for the annotation provision. As is described above, since
3 an annotation provided in advance can be copied for the
4 sharing portion, and the user need only add the annotation
5 for portions other than the sharing portion. As a result,
6 the workload required for the provision of the annotation can
7 be reduced. When the application of the annotation to the
8 sharing portion is impossible, the provision of the
9 annotation for the entire page is requested (step 85).
10 Thereafter, the same process as in the addition of the
11 annotation is performed for the page group having the
12 temporary layout ID (step 87). And a check is performed to
13 determine whether the above process has been performed for
14 [generally to all] the page groups in the layout sharing
15 groups (step 88). When the page groups have been processed,
16 this processing is terminated (step 89). But when [generally
17 all] the pages have not yet been processed, program control
18 returns to step 81 and the processing is repeated. When the
19 annotation is not applied for the entire sharing layout, the
20 inter-layout distance calculation expression is also
21 corrected (step 87). The processes shown in Figs. 8 to 10
22 are performed in order for [generally all] the page groups,
23 and the addition of annotations to the entire site is
24 completed.

25 As is described above, the information processing system or
26 method of this embodiment can simultaneously add an
27 annotation to or apply it to pages having the same or similar
28 layout. Further, when the same layout is used for one part
29 of the pages, the addition and the application of the

1 annotation to this sharing portion can also be simplified.
2 Thus, the efficiency of the user's operation to add an
3 annotation can be considerably increased. The operating
4 efficiency is especially improved for a site, such as a news
5 site or a database site, whereat the volume of the page files
6 carried is large, and the layouts employed for the pages tend
7 to be used in common.

8 When the user changes the determination of the similarity
9 that is automatically performed by the system, only the
10 distance calculation expression need be changed in the above
11 described manner, since the system automatically changes the
12 determination reference. Thus, the grouping accuracy can be
13 improved. As the determination reference is changed by the
14 user operation performed to provide an annotation, the user
15 need only provide an annotation for the operating efficiency
16 to be automatically improved. That is, as learning effect,
17 the reference for determining the layout group or the layout
18 sharing group is automatically changed by the user operation
19 that is performed. In this embodiment, an example for the
20 simultaneous provision of an annotation has been explained.
21 However, an annotation that has already been provided can be
22 used for the dynamic provision of an annotation for a page
23 file, and for transcoding, as follows.

24 Specifically, while a user is browsing an HTML document, an
25 annotation, such as "marking", is provided to a specific
26 position, and the system stores this information with the
27 layout data (layout tags and characteristic values) for a
28 pertinent page. During the browsing performed thereafter,

1 the user employs this layout data to perform transcoding,
2 such as division of a screen or the embedding of a link at a
3 marked position.

4 Further, when the browsing of a page having no annotation is
5 requested, the inter-page distance calculation module
6 calculates a distance between a requested page and a page for
7 which annotations have already been registered. As a result,
8 when the inter-page distance is smaller than a threshold
9 value, transcoding is performed using the annotation provided
10 for the nearest page, and the results are presented to the
11 user. When the user points at an annotation error, the
12 correction module for the function of distance calculation
13 changes the distance calculation expression. Further, the
14 user can add new annotation information, as needed. With
15 this method, since the user can add an annotation as needed
16 while browsing, instead of adding annotations for all the
17 pages in advance, the annotations can be added to the entire
18 site, step by step.

19 The invention has been specifically explained for an example
20 embodiment; however, the present invention is not limited to
21 this embodiment, and can be variously changed without
22 departing from the scope of the invention. For example, in
23 the above embodiment, to determine the similarity between the
24 page files, the method has been explained whereby the
25 distance between the pages or between the layout groups is
26 calculated by weighting the layout tags and characteristic
27 values. However, the method is not thereby limited, and a
28 tag skeleton method may be employed, or the similarity of the

1 images or the contents (text) of HTML documents may be
2 employed as a determination reference.

3 In addition, in this embodiment, the acquisition of the
4 layout sharing group and the application of an annotation to
5 a sharing layout using the layout sharing group need not be
6 requisite conditions for the present invention. In other
7 words, the present invention includes a case that is limited
8 to the acquisition of the layout groups and the application
9 of the annotation to the layout group. In this case, the
10 effects provided by the invention, such as the reduction in
11 the labor required for providing annotations, can be
12 obtained. Furthermore, in this invention, the condition for
13 correcting the calculation expression for the distances
14 between pages or layout groups need not be a requisite
15 condition. In this event, effects otherwise provided by the
16 invention can also be obtained.

17 In this embodiment, the similarities between the layouts of
18 HTML documents are employed to form groups. However, the
19 present invention can be extended to a determination of the
20 similarities between tags that are not related to the layout,
21 or the similarities of the contents of a document. In this
22 case, the similarities evidenced by HTML document structures
23 or the contents of documents can be determined, and this
24 determination can be employed for an analysis, for example,
25 of a site by a site manager, or for an analysis of a history
26 for the changing a page file at a site. Further, in the
27 example embodiment, an HTML file has been used as a page
28 file. However, the present invention can be applied for a

1 page file written in a markup language, such as XML

2 (Extensible Markup Language) or dynamic HTML.

3 Thus, this invention includes an operation for providing an
4 annotation for a page file can be efficiently performed. And
5 in addition, using the system of the invention, layout groups
6 or layout sharing groups can be more accurately formed.

7 The present invention can be realized in hardware, software,
8 or a combination of hardware and software. A visualization
9 tool according to the present invention can be realized in a
10 centralized fashion in one computer system, or in a
11 distributed fashion where different elements are spread
12 across several interconnected computer systems. Any kind of
13 computer system - or other apparatus adapted for carrying out
14 the methods and/or functions described herein - is suitable.

15 A typical combination of hardware and software could be a
16 general purpose computer system with a computer program that,
17 when being loaded and executed, controls the computer system
18 such that it carries out the methods described herein. The
19 present invention can also be embedded in a computer program
20 product, which comprises the features enabling the
21 implementation of the methods described herein, and which -
22 when loaded in a computer system - is able to carry out these
23 methods.

24 Computer program means or computer program in the present
25 context include any expression, in any language, code or
26 notation, of a set of instructions intended to cause a system
27 having an information processing capability to perform a
28 particular function either directly or after conversion to

1 another language, code or notation, and/or reproduction in a
2 different material form.

3 Thus the invention includes an article of manufacture which
4 comprises a computer usable medium having computer readable
5 program code means embodied therein for causing a function
6 described above. The computer readable program code means in
7 the article of manufacture comprises computer readable
8 program code means for causing a computer to effect the steps
9 of a method of this invention. Similarly, the present
10 invention may be implemented as a computer program product
11 comprising a computer usable medium having computer readable
12 program code means embodied therein for causing a a function
13 described above. The computer readable program code means in
14 the computer program product comprising computer readable
15 program code means for causing a computer to effect one or
16 more functions of this invention. Furthermore, the present
17 invention may be implemented as a program storage device
18 readable by machine, tangibly embodying a program of
19 instructions executable by the machine to perform method
20 steps for causing one or more functions of this invention.

21 It is noted that the foregoing has outlined some of the more
22 pertinent objects and embodiments of the present invention.
23 This invention may be used for many applications. Thus,
24 although the description is made for particular arrangements
25 and methods, the intent and concept of the invention is
26 suitable and applicable to other arrangements and
27 applications. It will be clear to those skilled in the art
28 that modifications to the disclosed embodiments can be

1 effected without departing from the spirit and scope of the
2 invention. The described embodiments ought to be construed
3 to be merely illustrative of some of the more prominent
4 features and applications of the invention. Other beneficial
5 results can be realized by applying the disclosed invention
6 in a different manner or modifying the invention in ways
7 known to those familiar with the art.